

# A Unified Monte-Carlo Jackknife for Small Area Estimation after Model Selection

JIMING JIANG

*University of California, Davis, U.S.A., jimjiang@ucdavis.edu*

P. LAHIRI

*University of Maryland, College Park, U.S.A.*

THUAN NGUYEN

*Oregon Health & Science University, Portland, U.S.A.*

We consider estimation of measure of uncertainty in small area estimation (SAE) when a procedure of model selection is involved prior to the estimation. A unified Monte-Carlo jackknife method, called McJack, is proposed for estimating the logarithm of the mean squared prediction error. We prove the second-order unbiasedness of McJack, and demonstrate the performance of McJack in assessing uncertainty in SAE after model selection through empirical investigations that include simulation studies and real-data analyses.

*Key Words.* Computer intensive, Jackknife, log-MSPE, measure of uncertainty, model selection, Monte-Carlo, second-order unbiasedness, small area estimation.

## 1 Introduction

Small area estimation (SAE) has become a very active area of statistical research and applications. Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced based on direct sampling from the population due to certain limitations of the available data. Examples of small areas include a geographical region (e.g., a state, county, municipality, etc.), a demographic group (e.g., a specific age  $\times$  sex  $\times$  race group), a demographic group within a geographic region, etc. See, for example, Rao and Molina (2015) for an updated, comprehensive account of various methods used in SAE. Statistical models, especially mixed effects models, have played key roles in improving small area estimates by borrowing strength from relevant sources. Therefore, it is not surprising that model selection in SAE has received considerable attention in recent liter-

ature. See, for example, Jiang, Nguyen and Rao (2010), Datta, Hall and Mandal (2011), Pfeiffermann (2013), Lahiri and Suntornchost (2014), and Rao and Molina (2015).

The errors from model selection are likely to affect the uncertainty measures in SAE estimates. To elaborate this point, let us consider a specific aspect of model selection—inclusion of small area specific random effects. Should one include area specific random effect in small area modeling? Such a component is a compromise between area specific fixed effects and no area effect and helps improving the properties of model-based estimators. For example, without such an area specific random effect, model-based estimator may not be design-consistent, which may result in model-based estimate for an area with large sample size to deviate significantly from the corresponding design-based estimate, especially if area specific auxiliary variables fail to capture variation across the areas. A decision to exclude small area specific random effect may be based on a significance test. But such a decision is anything but perfect and depends very much on the subjective choice of the prespecified level of significance. A reasonable uncertainty measure estimator must incorporate the impact of model selection. However, most of the uncertainty measure estimators, with the exception of Molina, Rao and Datta (2015), do not attempt to capture the variation due to the model choice and there is no analytical study to examine the important second-order unbiasedness property of any of these estimators, including that of Molina et al. (2015). In this paper, we propose a new uncertainty measure of any small area model-based estimator that incorporates errors due to model selection and a Monte-Carlo jackknife second-order unbiased estimator of the proposed uncertainty measure. We propose to use the logarithm of the mean squared prediction error (MSPE) as the uncertainty measure, where MSPE incorporates errors due to model selection. Our rationale behind using the log-MSPE comes from the way lack-of-fit measure of a typical model selection criterion is constructed. To elaborate on this point, consider the case of regression model selection with normal data. The well-known information criteria take the form of

$$n \log(\hat{\sigma}^2) + \lambda_n |M|, \quad (1)$$

where  $n$  is the sample size,  $\hat{\sigma}^2$  is the standard estimator of the error variance,  $\sigma^2$ ,  $|M|$  is the

dimension of the model,  $M$ , typically defined as the number of free parameters under  $M$ , and  $\lambda_n$  is a penalty function. Thus, in this case, the measure of lack-of-fit is proportional (under a fixed sample size) to the logarithm of a variance estimator. Note that, typically, the variance is of the same scale as the MSPE. Therefore, it is reasonable to consider the logarithm of the MSPE as a measure of uncertainty in SAE when a model selection procedure, such as an information criterion, is involved.

Besides the intuitive link to model selection, there are other advantages of using the log-MSPE as a measure of uncertainty. In the SAE literature, MSPE estimates have been routinely used in assessing an improvement of the empirical best linear unbiased predictor (EBLUP) over the direct estimator. For such a purpose, one can equivalently use the log-MSPE, and report the improvement in the log-scale. An advantage of log-MSPE over MSPE occurs when it is desirable to model uncertainty measure estimators. This is because one can reasonably assume normality of the error term when log-MSPE estimators are considered. Zimmerman *et al.* (1999) emphasized the need to model log-MSPE in the context of a geo-spatial application. Gershunskaya and Dorfman (2013) considered modeling of logarithm of variances in an application related to Current Employment Statistics survey. In a small area context, such a model can provide a guideline for making important decisions on the choice of different design factors (e.g., sample size, number of clusters) for a future survey in achieving, approximately, a certain desired level of log-MSPE of the proposed predictor for different small areas. Also, the model can be used for quickly producing uncertainty measures when it is time consuming to compute such measures when dealing with big data as well as computational complexity to meet a tight production deadline.

In terms of statistical inference, it is easier to carry out hypothesis testing when considering log-MSPE. For example, suppose that one wishes to compare  $\text{MSPE}_1$  with  $\text{MSPE}_2$ , which may correspond to two different methods of SAE. If one has second-order unbiased estimators of the log-MSPEs, say,  $\hat{l}_j$  for  $l_j = \log(\text{MSPE}_j)$ ,  $j = 1, 2$ , it is possible to construct a z-test, or t-test, by assuming (approximately) that  $\hat{l}_j = l_j + e_j$ ,  $j = 1, 2$ , where  $e_j$  is normal with mean zero and constant variance.

Finally, a desirable property for an MSPE estimator is that it needs to be positive. If

the property is combined with the second-order unbiasedness property, it turns out that it is very difficult to produce an estimator that has both of these properties. Typically, it is relatively easy to obtain a positive MSPE estimator that is first-order unbiased. To achieve the second-order unbiasedness, either analytical (e.g., Prasad and Rao 1990) or resampling (e.g., Jiang, Lahiri and Wan 2002, Hall and Maiti 2006) methods are used. However, with very few exceptions (Prasad and Rao 1990, Chen and Lahiri 2011), these techniques do not produce MSPE estimators that are guaranteed positive, in spite of achieving the second-order unbiasedness. To ensure that the MSPE estimator is positive, some modification of the (second-order unbiased) MSPE estimator is often made. For example, Hall and Maiti (2006) suggested the following strategy. Let  $\widehat{\text{MSPE}}_1$  and  $\widehat{\text{MSPE}}_2$  be two estimators of the same MSPE, for example, the former being an MSPE estimator with a additive bias-correction, and the latter one with a multiplicative bias-correction. Both MSPE estimators have some types of problems. For example,  $\widehat{\text{MSPE}}_1$  can take negative values, and  $\widehat{\text{MSPE}}_2$  can be unreliable (Hall and Maiti 2006). The idea is to combine the two estimators by letting  $\widehat{\text{MSPE}} = \widehat{\text{MSPE}}_1$  if something happens, and  $\widehat{\text{MSPE}} = \widehat{\text{MSPE}}_2$  otherwise. This strategy takes care of the positivity issue, but it does not necessarily preserve the second-order unbiasedness, even if  $\widehat{\text{MSPE}}_1$  and  $\widehat{\text{MSPE}}_2$  are both second-order unbiased. In fact, no rigorous proof has even been given that such a combined MSPE estimator is both positive and second-order unbiased. In contrast, there is no requirement that log-MSPE needs to be positive. Therefore, for log-MSPE, one can simply focus on the second-order unbiasedness of its estimator. Question is: How to obtain such an estimator?

In the context of MSPE estimation, a standard approach is Prasad-Rao (P-R) linearization (Prasad and Rao 1990). However, the approach is not feasible to handle our current problem, which is much more complicated. More specifically, we are interested in estimating the log-MSPE when the small area predictor is obtained after a model-selection procedure. The existing literature on inference after model selection has mainly focused on the case of independent observations (e.g., Rao and Wu 2001, sec. 12 and the references therein, Leeb 2009, Berk, Brown and Zhao 2010). In particular, the potential impact of model selection on MSPE has never been rigorously addressed in the SAE literature. Intu-

itively, there is an additional uncertainty involved in the model-selection process, that needs to be taken into account in the MSPE estimation. The P-R linearization method requires differentiability of the underlying operation. This usually holds for standard estimation and prediction procedures, but not for model selection. For example, the information criteria, such as AIC (Akaike 1973) and BIC (Schwarz 1978), or the fence methods (see Jiang 2014 for a review), select models from a discrete space of candidate models. Even the shrinkage methods (e.g., Tibshirani 1996, Fan and Li 2001) involve continuous but non-differentiable penalty functions, such as the  $L^1$  norm. See Müller, Scealy and Welsh (2013) for a review. Even if it is possible to develop a P-R type method, the derivation is tedious, and the final analytic expression is likely to be complicated. More importantly, errors often occur in the process of derivations as well as computer programming based on the lengthy expressions.

In this paper, we develop a unified jackknife approach that is assisted by Monte-Carlo simulations for the estimation of log-MSPE. As will be seen, the approach is applicable not just to the current problem of SAE after model selection, but to a much broader class of problems to obtain nearly unbiased estimators of quantities that can be obtained via Monte-Carlo simulation, if one knows the parameters that are involved. The method is especially attractive if the quantity of interest does not carry a constraint, such as non-negativity. This will be the case for the log-MSPE. Furthermore, the Monte-Carlo jackknife method, called *McJack*, is “one-formula-for-all”, which means that one needs not to re-derive the formula, as in P-R type methods, every time there is a new problem.

In the context of resampling methods, a well-known method is jackknife-after-bootstrap (JAB; Efron 1992). There are major differences between JAB and McJack. First, the objectives are different. The main purpose of JAB is to assess accuracy of the usual bootstrap estimates; while the objective of McJack is to estimate quantities of interest, such as measures of uncertainty for estimates based on the original data. Secondly, JAB works, for the most part, under the standard nonparametric bootstrap setting, to achieve efficient computation so that no additional bootstrap samples are needed; in other words, the JAB estimates are obtained from the original bootstrap samples. However, this is difficult to do under a parametric bootstrap setting. For example, although Efron (1992) has discussed JAB with

parametric bootstrap using the idea of importance sampling, the approach does not necessarily lead to a real gain in computation if the major computational burden is not due to sampling. On the other hand, standard nonparametric bootstrap procedures do not apply to SAE problems, in spite of some variations that have been developed. See, for example, Pfeiffermann (2013), for a review. Finally, McJack does not have to be associated with bootstrap—any kind of Monte-Carlo method can be used to assist the computation. For example, Jiang, Lahiri and Wan (2002; hereafter, JLW) discussed an example in which the Monte-Carlo method used to compute the MSPE is not considered as bootstrapping.

The rest of the paper is organized as follows. We begin by offering a critical review of JLW, which has had significant impact in SAE. We point out some undesirable features of JLW, and make two important observations that lead to McJack. The latter is described in Section 3 with a theoretical justification. Estimation of log-MSPE in SAE after model selection is illustrated using an example. In Section 4, we carry out simulation studies on performance of McJack, and compare it with alternative approaches. A real data application is considered in Section 5. We offer some discussion in Section 6. Proofs of the theorems are given in Section 7.

## 2 A brief review of JLW, and important observations

In the context of resampling methods for SAE, Jiang, Lahiri, and Wan (2002; hereafter, JLW) proposed a jackknife method for estimating the MSPE of empirical best predictor (EBP) when the parameters of interest are estimated by M-estimators. Let  $\xi$  denote a mixed effect, for example, a small area mean. Let  $\tilde{\xi}$  and  $\hat{\xi}$  denote the best predictor (BP), defined as conditional expectation of  $\xi$  given the data,  $y$ , and EBP of  $\xi$ , respectively. Then, one has the decomposition:

$$\text{MSPE}(\hat{\xi}) = \text{MSPE}(\tilde{\xi}) + E\{(\hat{\xi} - \tilde{\xi})^2\}, \quad (2)$$

where  $\text{MSPE}(\hat{\xi})$  is defined as  $E\{(\hat{\xi} - \xi)^2\}$  and  $\text{MSPE}(\tilde{\xi})$  is defined similarly. The idea of JLW is to jackknife the two terms on the right side of (2) separately. For the first term, the

authors assume that it is a function of  $\psi$ , a vector of parameters, that is,  $\text{MSPE}(\tilde{\xi}) = b(\psi)$ , which can be computed analytically. The parameter vector  $\psi$  is then estimated by an M-estimator, defined as the solution,  $\hat{\psi}$ , to a system of equations of the following form:

$$\sum_{i=1}^m f_i(\psi, y_i) + a(\psi) = 0. \quad (3)$$

In (3),  $y_i$  is the data vector from the  $i$ th cluster (e.g., small area), and the clusters are assumed to be independent;  $f_i(\cdot, \cdot)$  is a vector-valued function that satisfies  $E\{f_i(\psi, y_i)\} = 0$ ,  $1 \leq i \leq m$ , if  $\psi$  is the true parameter vector; and  $a(\cdot)$  corresponds to a penalizer, which in some cases is the zero vector. The delete- $j$  estimator,  $\hat{\psi}_{-j}$ , of  $\psi$  is defined as the solution to the following system of equations:

$$\sum_{i \neq j} f_i(\psi, y_i) + a_{-j}(\psi) = 0, \quad (4)$$

where  $a_{-j}(\cdot)$  has a similar interpretation. Given the M-estimators,  $b(\psi)$  is estimated by a plug-in estimator, minus a jackknife bias correction, that is,

$$b(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^m \{b(\hat{\psi}_{-j}) - b(\hat{\psi})\}. \quad (5)$$

As for the second term on the right side of (2), it is estimated by a jackknife variance-type estimator that has the following expression:

$$\frac{m-1}{m} \sum_{j=1}^m (\hat{\xi}_{-j} - \hat{\xi})^2, \quad (6)$$

where  $\hat{\xi}_{-j}$  is a delete- $j$  version of  $\hat{\xi}$ , the EBP, defined in a certain way, which is not important for the current paper. JLW showed that, when the two terms, (5) and (6), are put together, the combined jackknife estimator of the MSPE of EBP is second-order unbiased. The work has had a significant impact in SAE, especially in the literature of resampling methods in SAE (e.g., Hall and Maiti 2006, Lohr and Rao 2009, Pfeiffermann 2013, Rao and Molina 2015). On the other hand, we note the following undesirable features of JLW:

(a) JLW requires analytical computation of  $b(\psi)$ . More specifically, JLW assumes posterior



linearity, under which  $b(\psi)$  has an analytic expression.

(c) JLW does not incorporate errors from model selection. In particular, the proof for the second-order unbiased property of (6) fails if a model selection procedure is involved prior to obtaining the EBP, such as in Datta *et al.* (2011).

(c) JLW does not ensure a strictly positive MSPE estimator, in spite of its second-order unbiasedness. See our discussion in Section 1 (5th paragraph).

As far as this paper is concerned, what is most important is not the full JLW theory, but rather an intermediate result. In obtaining their theory, JLW showed, in particular, that (5) is a second-order unbiased estimator of  $b(\psi)$ , if the penalizers  $a, a_{-j}, 1 \leq j \leq m$  in (3) and (4) satisfy certain mild conditions. In particular, those conditions are satisfied if the penalizers are zero (vectors), in which case the M-estimating equations are unbiased. Having given the proof of the result, we realize the following two facts, both are critically important to the idea of the current paper.

(I) The fact that  $b(\psi)$  is an MSPE is not used anywhere in the proof. In other words, as long as  $b(\cdot)$  is a sufficiently smooth function, and  $\psi$  is estimated by the M-estimators, the second-order unbiased estimation of  $b(\psi)$  by (5) holds. In particular,  $b(\psi)$  can be  $\log(\text{MSPE})$ , which is of primary interest here.

(II) More importantly,  $b(\psi)$  does not have to have an analytic expression, as long as one knows how to compute it. An analytic expression would be nice, but, in the new era, the computation is typically done by a computer, perhaps, a high-powered one. In particular, suppose that, given  $\psi$ ,  $b(\psi)$  can be approximated by a Monte-Carlo method to an arbitrary degree of accuracy. Then, one can write a computer program, based on the Monte-Carlo, to compute  $b(\cdot)$  as a function. Given this “computer-powered” function, all one needs to do is to plug the M-estimators,  $\hat{\psi}, \hat{\psi}_{-j}, 1 \leq j \leq m$ , into this function to obtain the second-order unbiased estimator of  $b(\psi)$ .

The importance of the above observations is that they apply to virtually any kind of situation, not just the EBP. In particular, the predictor,  $\hat{\xi}$ , can be much more complicated than the EBP, such as an EBP obtained following a model-selection procedure. Also, the decomposition (2), the posterior linearity assumption, and (6) are altogether not needed to



apply these observations. In the next section, we propose a new method based on the two important observations that addresses all of the undesirable features of JLW noted above. Other complicated situations, to which our idea may apply, include (i) regression inference after variable selection (e.g., Leeb 2009); (ii) mixed model prediction with non-normal random effect distribution (e.g., Lahiri and Rao 1995); and (iii) shrinkage estimation/selection with data-driven choice of regularization parameter (e.g., Pang, Lin and Jiang 2015).

### 3 Monte-Carlo jackknife

We first illustrate the method using an example of EBLUP under a Fay-Herriot model, where the BIC (Schwarz 1978) is used to select the fixed covariates as well as whether to include the area-specific random effects. The model can be expressed in a way more convenient for the model selection problem:

$$y_i = x_i' \beta + \sqrt{A} \xi_i + e_i, \quad (7)$$

$i = 1, \dots, m$ , where the components of  $x_i$  are to be selected from a set of candidate covariates;  $\xi_i \sim N(0, 1)$ ; if  $A > 0$ , the random effects are included in the model; if  $A = 0$ , the random effects are excluded from the model;  $e_i \sim N(0, D_i)$ , where  $D_i, 1 \leq i \leq m$  are known; and the  $\xi_i$ 's and  $e_i$ 's are independent. Note that there have been further considerations regarding the choice of the random effects; see, for example, Datta *et al.* (2011), but here we focus on a simpler situation. Let  $M_f$  denote a full model, under which  $x_i$  is the vector that includes all of the candidate covariates, and  $A \geq 0$ . Denote the  $x_i$  under  $M_f$  by  $x_{f,i}$ , and the corresponding  $\beta$  by  $\beta_f$ . Let  $\psi = (\beta_f', A)'$ . It is easy to see that  $M_f$  is, at least, a correct model, which means that (7) holds with  $x_i$  replaced by  $x_{f,i}$ ,  $\beta$  replaced by  $\beta_f$ , and the range of  $A$  being  $[0, \infty)$ . Of course, some of the components of  $\beta_f$  may be zero, in case that the full model can be simplified, and the true  $A$  may be zero—these are the reasons for the model selection. But this does not change the fact  $M_f$  is a correct model. In particular, the true small-area mean,  $\theta_i$ , can be expressed as

$$\theta_i = x_{f,i}' \beta_f + \sqrt{A} \xi_i. \quad (8)$$

On the other hand, under a candidate model,  $M$ , which corresponds to (7), the EBLUP of  $\theta_i$  can be expressed as

$$\tilde{\theta}_i = \frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i} x'_i \hat{\beta}, \quad (9)$$

where  $\hat{\beta} = \{\sum_{i=1}^m (\hat{A} + D_i)^{-1} x_i x'_i\}^{-1} \sum_{i=1}^m (\hat{A} + D_i)^{-1} x_i y_i$ , and  $\hat{A}$  is a consistent estimator of  $A$  obtained using a certain method (e.g., P-R, ML, REML; see Rao and Molina 2015). The BIC procedure chooses the model,  $M$ , by minimizing

$$\text{BIC}(M) = -2\hat{l} + |M| \log(m), \quad (10)$$

where  $\hat{l}$  is the maximized log-likelihood under  $M$ ;  $|M| = \dim(\beta) + 1$  if  $M$  includes the random effects, and  $|M| = \dim(\beta)$  if  $M$  excludes the random effects. Here, for simplicity, we assume that  $X = (x'_i)_{1 \leq i \leq m}$  is full rank under any  $M$ . Let the minimizer of (10) be  $\hat{M}$ . We then compute the EBLUP (9) under  $M = \hat{M}$ , that is,

$$\hat{\theta}_i = \frac{\hat{A}_{\hat{M}}}{\hat{A}_{\hat{M}} + D_i} y_i + \frac{D_i}{\hat{A}_{\hat{M}} + D_i} x'_{\hat{M},i} \hat{\beta}_{\hat{M}}, \quad (11)$$

where  $\hat{\beta}_{\hat{M}}$  and  $\hat{A}_{\hat{M}}$  are the  $\hat{\beta}$  and  $\hat{A}$  obtained under  $\hat{M}$ . The MSPE of interest is

$$\text{MSPE}(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i)^2, \quad (12)$$

where  $\theta_i$  is given by (8). It is clear that the joint distribution of  $(\theta_i, y_i)$ ,  $1 \leq i \leq m$  depends only on  $\psi = (\beta'_f, A)$ . Thus, (12) is a function of  $\psi$  and so is its logarithm. Let

$$b(\psi) = \log\{\text{MSPE}(\hat{\theta}_i)\}. \quad (13)$$

Given  $\psi$ , for the  $k$ th Monte-Carlo simulation, one first generates  $\theta_i$  by (8) with  $\xi_i$  replaced by  $\xi_i^{(k)}$ ,  $1 \leq i \leq m$ , generated independently from  $N(0, 1)$ . Denote the generated  $\theta_i$  by  $\theta_i^{(k)}$ . Next, let  $y_i^{(k)} = \theta_i^{(k)} + e_i^{(k)}$ ,  $1 \leq i \leq m$ , where  $e_i^{(k)} \sim N(0, D_i)$ ,  $1 \leq i \leq m$ , generated independently and independent with  $\xi_i^{(k)}$ 's. The Monte-Carlo approximation to  $b(\psi)$  is

$$\tilde{b}(\psi) = \log \left[ \frac{1}{K} \sum_{k=1}^K \left\{ \hat{\theta}_i^{(k)} - \theta_i^{(k)} \right\}^2 \right], \quad (14)$$

where  $\hat{\theta}_i^{(k)}$  is obtained the same way as the  $\hat{\theta}_i$  of (11) except with  $y_i$  replaced by  $y_i^{(k)}$ ,  $1 \leq i \leq m$ . Write the above procedure as a function, say,  $\tilde{b}(\psi) = \mathbf{mcjack}(\psi)$ , that computes (14) for every given  $\psi$ . Now suppose that  $\hat{\psi}$  is an M-estimator of  $\psi$ . For example,  $\hat{A}$  is the P-R estimator (Prasad and Rao 1990; truncated at zero if the expression turns out to be negative), and  $\hat{\beta}_f$  is given below (9) with  $x_i = x_{f,i}$ ,  $1 \leq i \leq m$ . Let  $\hat{\psi}_{-j}$  be the delete- $j$  version of  $\hat{\psi}$ . The McJack estimator of (13) is then given by

$$\widehat{b(\psi)} = \tilde{b}(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^m \{\tilde{b}(\hat{\psi}_{-j}) - \tilde{b}(\hat{\psi})\}. \quad (15)$$

Although the above illustration is based on the Fay-Herriot model, its general principle, namely, (12)–(15), applies to much broader cases. Using the result of JLW, we can justify the second-order unbiasedness of McJack under the general framework. The justification also takes into account effect of the Monte-Carlo errors. First note that, to establish a rigorous result about the unbiasedness, we need to make sure that the expectations of  $\tilde{b}(\hat{\psi}_{-j})$ ,  $0 \leq j \leq m$  exist. To avoid complicated technical conditions, we regularize these estimators (e.g., Jiang *et al.* 2002, Das *et al.* 2004). Let  $\tilde{s}(\psi) = \exp\{\tilde{b}(\psi)\}$ , and define

$$\hat{s}(\psi) = \begin{cases} e^{-\lambda m^\rho}, & \text{if } \tilde{s}(\psi) < e^{-\lambda m^\rho}, \\ \tilde{s}(\psi), & \text{if } e^{-\lambda m^\rho} \leq \tilde{s}(\psi) \leq e^{\lambda m^\rho}, \\ e^{\lambda m^\rho}, & \text{if } \tilde{s}(\psi) > e^{\lambda m^\rho}, \end{cases}$$

and  $\hat{b}(\psi) = \log\{\hat{s}(\psi)\}$ , where  $\lambda, \rho$  are given positive numbers. Let  $s(\psi)$  denote MSPE( $\hat{\theta}_i$ ) when  $\psi$  is the true parameter vector. We truncate  $s(\cdot)$  the same way as  $\tilde{s}(\cdot)$ , and let  $b(\psi) = \log\{s(\psi)\}$ . For notation convenience, write  $\hat{\psi}_{-0} = \hat{\psi}$ . Also, let  $F_{-0}(\psi)$ ,  $F_{-j}(\psi)$  denote the left sides of (3) and (4), respectively. The M-estimators,  $\hat{\psi}_{-j}$ ,  $0 \leq j \leq m$  are said to be consistent uniformly (c.u.) at rate  $m^{-d}$  if, for any  $\delta > 0$ , there is a constant  $c_\delta$  such that

$$P(A_{j,\delta}^c) \leq c_\delta m^{-d}, \quad 0 \leq j \leq m,$$

where  $A_{j,\delta}$  is the event that  $F_{-j}(\hat{\psi}_{-j}) = 0$  and  $|\hat{\psi}_{-j} - \psi| \leq \delta$ , with  $\psi$  being the true parameter vector. Also, write  $f_i = f_i(\psi, y_i)$ ,  $g_i = \partial f_i / \partial \psi'$ ,  $h_{i,k} = \partial^2 f_{i,k} / \partial \psi \partial \psi'$ , where

$f_{i,k}$  is the  $k$ th component of  $f_i$ . Furthermore, for any function  $f$  of  $\psi$ , define

$$\|\Delta^3 f\|_w = \max_{1 \leq s, t, u \leq r} \sup_{|\tilde{\psi} - \psi| \leq w} \left| \frac{\partial^3 f(\tilde{\psi})}{\partial \psi_s \partial \psi_t \partial \psi_u} \right|,$$

where  $\psi$  is the true parameter vector, and  $r = \dim(\psi)$ . A similar definition is extended to  $\|\Delta^4 f\|_w$ . The spectral norm of a matrix,  $B$ , is defined as  $\|B\| = \sqrt{\lambda_{\max}(B'B)}$ , where  $\lambda_{\max}$  denotes the largest eigenvalue. Also write  $\Delta_j = a - a_{-j}$ , where  $a, a_{-j}$  are the functions of  $\psi$  that appear in (3) and (4), respectively. We shall consider estimation of log-MSPE of  $\hat{\theta}_i$ , a predictor of  $\theta_i$  after model selection, for a fixed  $i$ . Furthermore, we assume that the Monte-Carlo samples, under  $\psi$ , are generated by first generating some standard [e.g.,  $N(0, 1)$ ] random variables and then plugging  $\psi$ . For example, under the full Fay-Herriot model of (7),  $y_i$  is generated by first generating the  $\xi_i$ 's and  $\eta_i$ 's, which are independent  $N(0, 1)$ , and then letting  $y_i = x'_{f,i}\beta_f + \sqrt{A}\xi_i + \sqrt{D_i}\eta_i$ , with  $\psi = (\beta'_f, A)'$ . Let  $\xi$  denote the vector of the standard random variables. We first make the following general assumptions.

A1. There are  $d > 2$  and  $w > 0$  such that the  $2d$ th moments of  $|f_i|$ ,  $\|g_i\|$ ,  $\|h_{i,k}\|$ ,  $\|\Delta^3 f_{i,k}\|_w$ ,  $1 \leq i \leq m$ ,  $1 \leq k \leq r$  are bounded for some  $d > 2 + \rho$ .

A2. For the same  $d$  and  $w$  in A1,  $a_{-j}$  and its up to third order partial derivatives,  $0 \leq j \leq m$ , as well as  $\Delta_j$ ,  $1 \leq j \leq m$ , all evaluated at  $\tilde{\psi}$ , are bounded uniformly for  $|\tilde{\psi} - \psi| \leq w$ , where  $\psi$  is the true parameter vector, and  $m^\tau(|\Delta_j| \vee \|\partial \Delta_j / \partial \psi\|)$ ,  $1 \leq j \leq m$ , evaluated at  $\psi$ , are bounded, where  $\tau = (d - 2)/(2d + 1)$ .

A3. The log-MSPE function  $b(\cdot)$  of (13) is four-times continuously differentiable, and, for the same  $w$  in A1,  $\|\Delta^4 b\|_w$  is bounded.

A4.  $\limsup_{m \rightarrow \infty} \|\{E(\bar{g})\}^{-1}\| < \infty$ , where  $\bar{g} = m^{-1} \sum_{j=1}^m g_j$ , evaluated at the true  $\psi$ .

A5.  $\hat{\psi}_{-j}$ ,  $0 \leq j \leq m$  are c.u. at rate  $m^{-d}$  for the same  $d$  in A1.

A6.  $\sum_{j=1}^m \Delta_j = O(m^{-\nu})$  for some  $\nu > 0$ .

Recall the way that the Monte-Carlo samples are generated specified above A1. Under this assumption,  $\theta_i^{(k)}, \hat{\theta}_i^{(k)}$ ,  $1 \leq k \leq K$ , generated under  $\tilde{\psi}$ , are functions of  $\tilde{\psi}$  and  $\xi$ . The additional assumptions below are regarding the Monte-Carlo sampling.

A7.  $\xi$  is independent with the data,  $y$ .

A8. Let  $\psi$  be the true parameter vector, and  $w$  be the same as in A1. There are constants  $0 < c_1 < c_2$  such that  $c_1 \leq s(\tilde{\psi}) \leq c_2$  for  $|\tilde{\psi} - \psi| \leq w$ , and random variables  $G_k$ ,  $1 \leq k \leq K$ , which do not depend on  $\tilde{\psi}$ , such that  $|\hat{\theta}_i^{(k)} - \theta_i^{(k)}| \leq G_k$  and  $E(G_k^q)$  are bounded for some  $q \geq 2\{2 + (\rho \vee 1)\}$ .

A9.  $m^2/K \rightarrow 0$ , as  $m \rightarrow \infty$ .

**Theorem 1.** Suppose that A1–A9 hold. Let  $\widehat{b(\psi)}$  denote (15) with  $\tilde{b}$  replaced by  $\hat{b}$ . Then, we have  $E\{\widehat{b(\psi)} - b(\psi)\} = o(m^{-1})$ , where  $\psi$  is the true  $\psi$  [hence  $b(\psi)$  is the true log-MSPE], and  $E$  is with respect to both  $y$  and  $\xi$ .

The next result focuses on the special case of Fay-Herriot model.

**Theorem 2.** Suppose that the true  $A > 0$ , and there are positive constants  $0 < c_1 < c_2$  such that  $c_1 \leq |x_{f,i}| \leq c_2$ ,  $c_1 \leq D_i \leq c_2$ ,  $1 \leq i \leq m$ . Furthermore, suppose that

$$\limsup_{m \rightarrow \infty} \lambda_{\min} \left( \frac{1}{m} \sum_{i=1}^m x_{f,i} x'_{f,i} \right) > 0, \quad (16)$$

and A9 holds. Then, the conclusion of Theorem 1 holds.

The proofs of Theorem 1 and Theorem 2 are given in Section 7.

## 4 Numerical demonstration and simulation study

### 4.1 A simple demonstration

To begin with, let us consider a very simple situation, which may be viewed as a special case of the Fay-Herriot model,

$$y_i = x'_i \beta + v_i + e_i, \quad i = 1, \dots, m, \quad (17)$$

where the components of  $x_i$  consist of an intercept, a group indicator,  $x_{1,i}$ , which is 0 if  $1 \leq i \leq m_1 = m/2$ , and 1 if  $m_1 + 1 \leq i \leq m$ , and potentially a third component,  $x_{2,i}$ , which is generated from the  $N(0, 1)$  distribution, and fixed throughout the simulation. There are two candidate models: Model 1, which includes  $x_{2,i}$ , and Model 2: which does not include  $x_{2,i}$ . The model selection is carried out by BIC (Schwarz 1978).

For this demonstration, we consider a special case that the variance of the random effects,  $v_i$ , is known to be zero, that is,  $A = 0$ . There have been considerations of such situations in SAE (e.g., Datta *et al.* 2011). The variance of  $e_i$ ,  $D_i$ , is equal to 1 for  $1 \leq i \leq m_1$ , and  $a$  for  $m_1 + 1 \leq i \leq m$ , where the value of  $a$  is either 4 or 16. Because  $A = 0$ , the small area mean,  $\theta_i$ , under a given model, is equal to  $x'_i\beta$ . The corresponding EBLUP is  $\hat{\theta}_i = x'_i\hat{\beta}$ , where  $\hat{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}y$ , with  $X = (x'_i)_{1 \leq i \leq m}$  and  $D = \text{diag}(D_i, 1 \leq i \leq m)$ , is the best linear unbiased estimator (BLUE) of  $\beta$  (e.g., Jiang 2007, sec. 2.3), under the given model. Due to the unbiasedness of the BLUE, the MSPE of the EBLUP is equal to its variance, that is,

$$\text{MSPE}(\hat{\theta}_i) = \text{var}(\hat{\theta}_i) = x'_i(X'D^{-1}X)^{-1}x_i, \quad 1 \leq i \leq m, \quad (18)$$

which are known under the given model. Now suppose that the EBLUP is obtained based on the model selected by the BIC. A naive estimator of the MSPE of  $\hat{\theta}_i$ , which ignores model selection, would be (18) computed under the selected model. The naive estimator of the log-MSPE is the logarithm of the naive MSPE estimator. We compare this estimator with two competitors. The first is what we call bootstrap MSPE estimator, which corresponds to the first term in (15), that is, without the jackknife bias correction, where  $b(\cdot)$  is the log-MSPE function. The second is the McJack estimator given by (15). The bootstrap and McJack estimators are computed based on  $K = 1000$  Monte-Carlo samples.

A series of simulation studies were carried out with  $m = 20$  and  $\beta_0 = \beta_1 = 1$ , where  $\beta_0$  is the intercept and  $\beta_1$  the slope of  $x_{1,i}$ , and under two different true underlying models. In the first scenario, Model 1 is the true underlying model with the slope of  $x_{2,i}$ ,  $\beta_2 = 0.5$ . In the second scenario, Model 2 is the true underlying model (i.e.,  $\beta_2 = 0$ ). We present the simulated percentage relative bias (%RB), based on  $N_{\text{sim}} = 1000$  simulation runs, in Figures 2 and 3, where, for a given area, the %RB is defined as

$$\%RB = \left[ \frac{E\{\log(\widehat{\text{MSPE}})\} - \log(\text{MSPE})}{|\log(\text{MSPE})|} \right] \times 100\%, \quad (19)$$

MSPE is the true MSPE based on the simulations, and  $E\{\log(\widehat{\text{MSPE}})\}$  is the mean of the estimated log-MSPE based on the simulations. It is seen that the naive estimator signif-

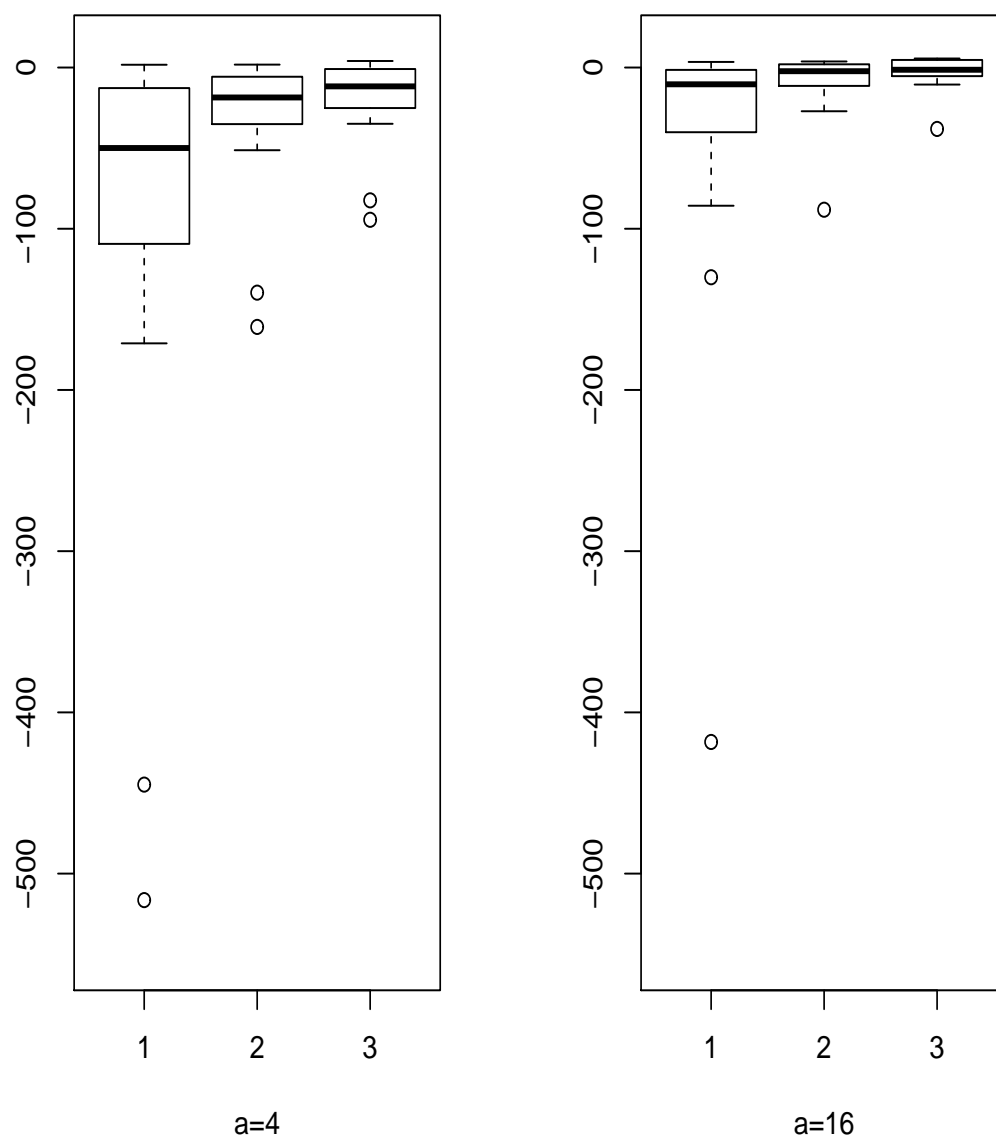


Figure 1: Boxplots of %RB when Model 1 is the true model. In each plot, from left to right: 1–Naive estimator, 2–bootstrap estimator, and 3–McJack estimator, of log-MSPE.



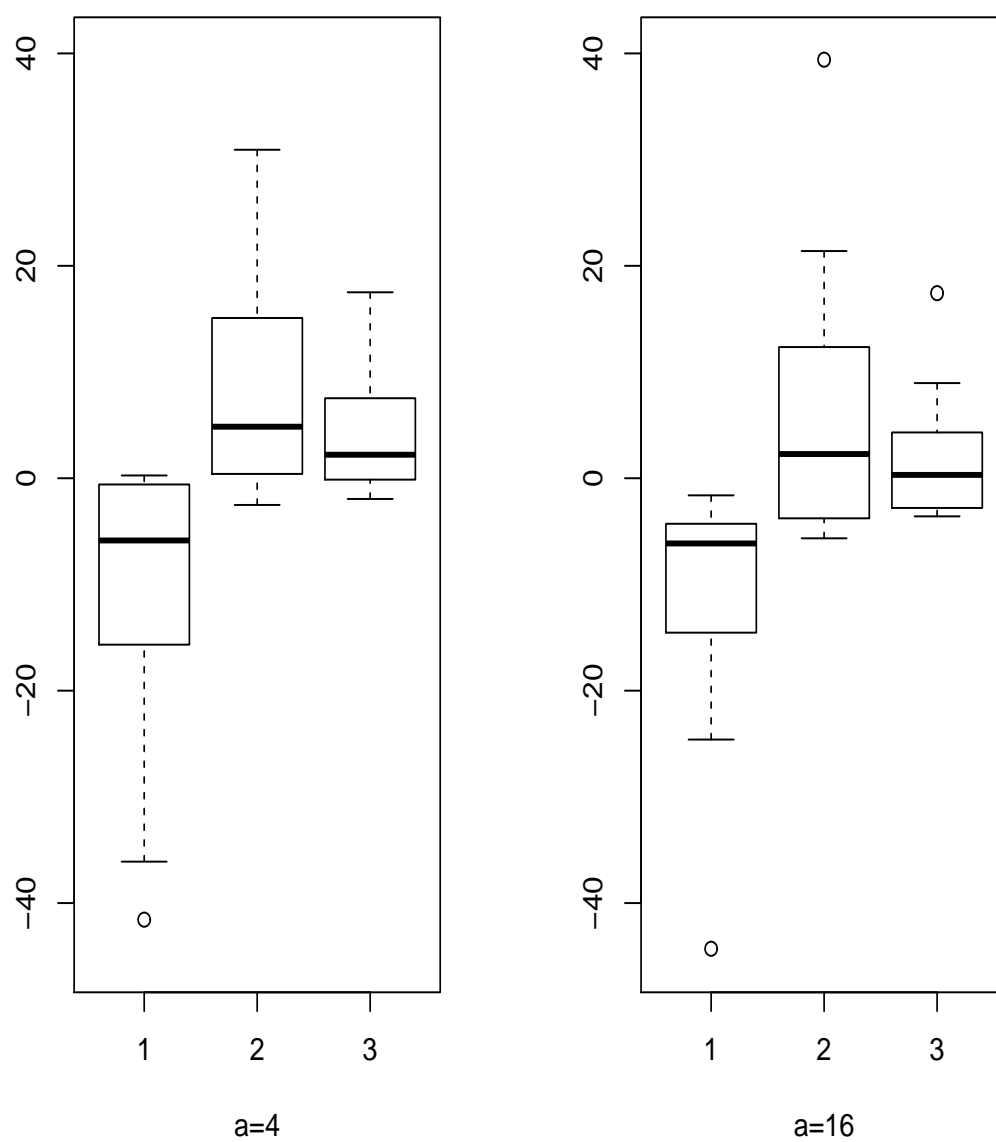


Figure 2: Boxplots of %RB when Model 2 is the true model. In each plot, from left to right: 1–Naive estimator, 2–bootstrap estimator, and 3–McJack estimator, of log-MSPE.

icantly under estimate the log-MSPE; in fact, when Model 1 is the true model, the %RB for one of the areas is 516% in the case of  $a = 4$ , and there is a similar case in the case of  $a = 16$ . More specifically, there are some interesting trend observed. Namely, when the true model is Model 1, all of the methods seem to under-estimate the log-MSPE, but the bootstrap and McJack estimators are doing much better, with McJack offering significant improvement over the bootstrap. On the other hand, when the true model is Model 2, the naive estimator again under-estimate the log-MSPE, but the bootstrap and McJack estimators seem to over-estimate the log-MSPE, with McJack significantly improving the bootstrap. The amount of underestimation by the naive estimator is less dramatic when Model 2 is the true model compared to when Model 1 is the true model. One explanation is that the BIC is known to have the tendency to over-penalize larger models. This would have bigger impact when Model 1 is the true model, which is the full model. In other words, there is a higher chance of model misspecification by the BIC, which impacts the log-MSPE estimation. To have a closer look at the numbers, we present one set of the detailed results in Table 2.

## 4.2 Testing the presence of random effects in a Fay-Herriot model

Datta *et al.* (2011) proposed a method of model selection by testing for the presence of the area-specific random effects,  $v_i = \sqrt{A}\xi_i$ , in the Fay-Herriot model (7). This is equivalent to testing the null hypothesis  $H_0 : A = 0$ . The test statistic,  $T = \sum_{i=1}^m D_i^{-1}(y_i - x_i'\hat{\beta})^2$ , where  $\hat{\beta}$  is the same as in Subsection 3.1, has a  $\chi_{m-p}^2$  distribution, with  $p = \text{rank}(X)$ , under  $H_0$ . If  $H_0$  is rejected, the EBLUP is used to estimate the small area mean  $\theta_i$ , where in this simulation  $A$  is estimated by the P-R estimator, and the corresponding MSPE estimator is the P-R MSPE estimator; if  $H_0$  is accepted, the estimator  $\hat{\theta}_i = x_i'\hat{\beta}$  is used to estimate  $\theta_i$ , and the corresponding MSPE is given by (18). Thus, if the level of significance is chosen as 0.05, the proposed MSPE estimator, denoted by DHM, is the P-R MSPE estimator if  $T > \chi_{m-p}^2(0.05)$ , and (18) if  $T \leq \chi_{m-p}^2(0.05)$ .

We run a simulation study to compare the performance of McJack with DHM. The

Table 1: **Log-MSPE estimation:** Model 2 is True Model;  $a = 4$ ; %RB in ( )s.

Area	True log-MSPE	E(Naive Est.)	E(Bootstrap Est.)	E(McJack Est.)
1	-1.98	-2.26 (-14.0)	-1.79 (9.6)	-1.91 (3.3)
2	-1.62	-2.21 (-36.1)	-1.22 (25.0)	-1.41 (12.8)
3	-2.07	-2.27 (-9.8)	-1.95 (5.6)	-2.01 (2.9)
4	-2.20	-2.30 (-4.3)	-2.26 (-2.5)	-2.25 (-1.9)
5	-1.70	-2.22 (-30.4)	-1.33 (21.7)	-1.52 (10.7)
6	-2.05	-2.27 (-10.8)	-1.91 (6.7)	-1.97 (3.7)
7	-2.14	-2.29 (-6.9)	-2.11 (1.5)	-2.16 (-1.0)
8	-1.55	-2.20 (-41.6)	-1.11 (28.4)	-1.28 (17.5)
9	-2.19	-2.30 (-4.8)	-2.23 (-1.6)	-2.22 (-1.4)
10	-2.06	-2.27 (-10.2)	-1.94 (6.0)	-2.00 (3.2)
11	-0.91	-0.92 (-0.5)	-0.91 (0.6)	-0.91 (-0.0)
12	-0.91	-0.92 (-0.1)	-0.91 (0.2)	-0.92 (-0.1)
13	-0.76	-0.89 (-17.4)	-0.61 (19.8)	-0.69 (9.5)
14	-0.87	-0.90 (-3.7)	-0.78 (10.4)	-0.82 (5.6)
15	-0.92	-0.92 (0.3)	-0.92 (0.1)	-0.92 (-0.1)
16	-0.92	-0.92 (0.1)	-0.92 (0.1)	-0.92 (-0.1)
17	-0.74	-0.88 (-18.1)	-0.52 (30.9)	-0.62 (17.2)
18	-0.92	-0.91 (0.1)	-0.90 (2.1)	-0.91 (1.1)
19	-0.91	-0.92 (-0.6)	-0.90 (0.7)	-0.91 (-0.0)
20	-0.88	-0.91 (-3.6)	-0.84 (4.1)	-0.87 (1.5)

simulation is under the full model considered in the previous subsection (hence  $p = 3$ ), and three different true values of  $A$ :  $A = 0$ ,  $A = 0.5$ , and  $A = 1$ . The boxplots of %RB for these three cases are presented in Figure 4, with the detailed numbers for DHM and McJack given in Table 3. It is seen that DHM works better for the case  $A = 0$ , which is not surprising because, under the null hypothesis, the DHM MSPE estimator is “right” 95% of times. On the other hand, McJack works significantly better in those two cases of nonzero  $A$ . Simple simulations show that, in the latter cases, the probability of rejecting the null hypothesis is about 0.26 when  $A = 0.5$ , and 0.44 when  $A = 1$ . The worst scenario seems to be the case where  $A$  is not zero but closer to zero ( $A = 0.5$ ). There are a few “blown-up” cases under this scenario where the %RB exceeds 1000% for DHM. It is also obvious that McJack improves bootstrap in every case.

Another simulated example, in which the model selection is carried out via a generalized information criterion (GIC) before the SAE, is also considered. The details are deferred to Supplementary Material due to the space limit.

## 5 A real data example

Morris and Christiansen (1995) presented a data set involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27 month period (see Table 5). The  $y_i$ 's are graft failure rates for kidney transplant operations, that is,  $y_i = \text{number of graft failures}/n_i$ , where  $n_i$  is the number of kidney transplants at hospital  $i$  during the period of interest. The variance for the graft failure rate,  $D_i$ , is approximated by  $(0.2)(0.8)/n_i$ , where 0.2 is the observed failure rate for all of the hospitals. Thus,  $D_i$  is assumed known. In addition, a severity index,  $s_i$ , is available for each hospital, which is the average fraction of females, blacks, children and extremely ill kidney recipients at hospital  $i$ . Ganesh (2009) proposed a Fay-Herriot model for the graft failure rates, which is (2) with  $x_i'\beta = \beta_0 + \beta_1 s_i$ . Jiang *et al.* (2010) suggests that, in a way, the optimal model for this data is a cubic model, that is, (2) with  $x_i'\beta = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 s_i^3$ , which is also used in Datta *et al.* (2011).

We analyze the data under the latter model for the mean function but with selection of

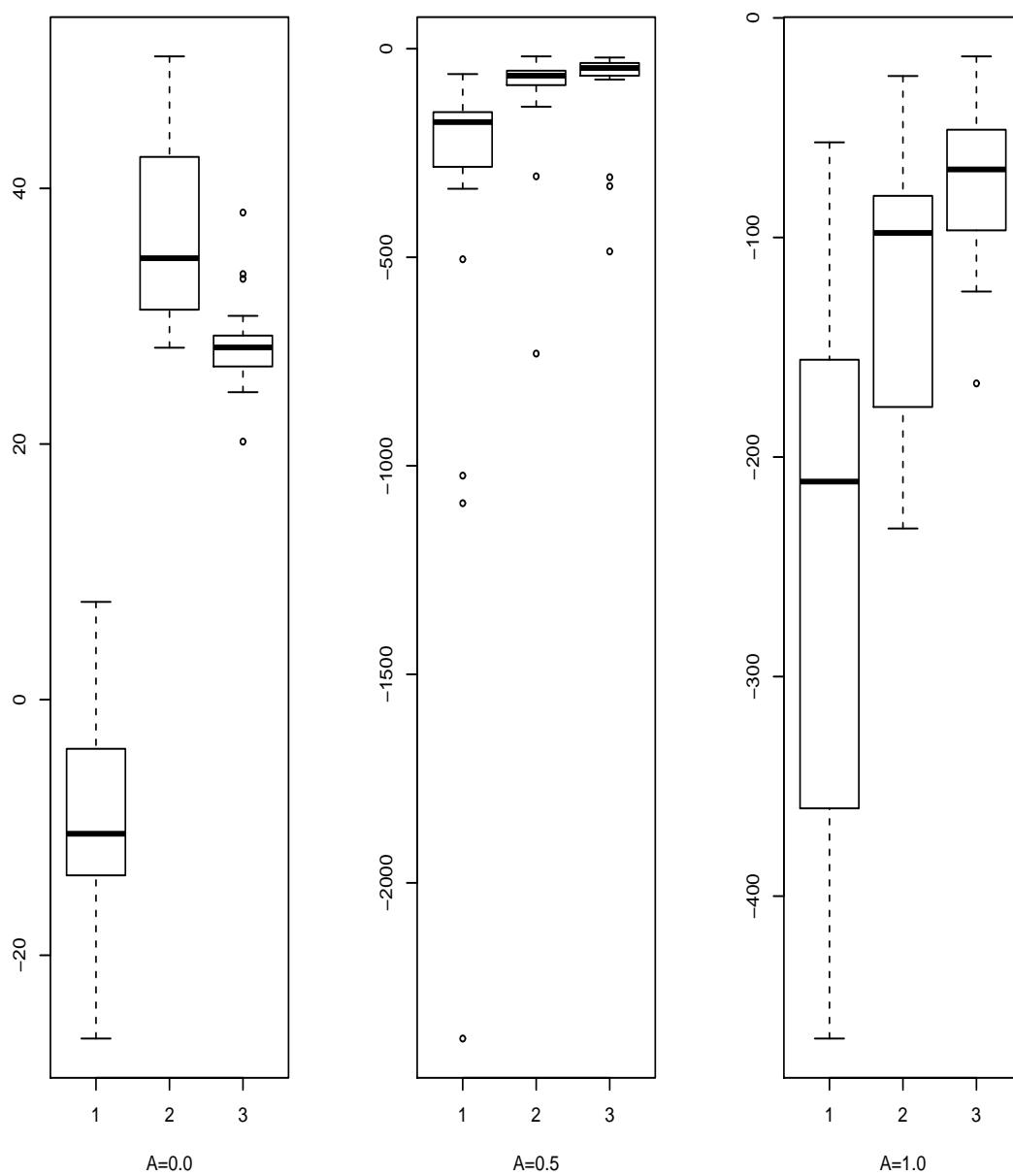


Figure 3: Boxplots of %RB. In each plot, from left to right: 1-DHM, 2-bootstrap, 3-McJack. Scales are different due to the huge difference in range.

Table 2: **DHM vs McJack in %RB**

Area	$A = 0.0$		$A = 0.5$		$A = 1.0$	
	DHM	McJack	DHM	McJack	DHM	McJack
1	-13.6	26.2	-216.8	-59.8	-342.0	-99.5
2	0.3	26.0	-131.0	-45.8	-343.8	-72.9
3	-3.9	27.5	-107.4	-21.1	-135.8	-30.3
4	1.1	28.2	-158.0	-37.7	-362.9	-77.6
5	-8.1	24.9	-178.9	-36.9	-191.3	-59.6
6	-6.0	30.0	-180.3	-50.5	-375.4	-166.5
7	-3.1	24.1	-210.0	-51.5	-395.5	-124.6
8	7.6	27.6	-135.3	-33.1	-464.9	-123.0
9	-10.9	27.4	-149.4	-43.0	-357.2	-108.1
10	-3.8	28.6	-163.1	-31.4	-362.8	-94.0
11	-26.5	20.2	-60.8	-21.8	-220.3	-39.4
12	-10.7	33.3	-2373.3	-486.1	-210.4	-76.5
13	-13.9	27.5	-504.8	-74.0	-94.5	-18.4
14	-10.3	32.9	-173.1	-35.3	-188.2	-64.1
15	-17.5	25.7	-1023.6	-329.5	-163.0	-58.1
16	-4.4	38.1	-154.6	-46.6	-211.9	-65.2
17	-12.1	28.4	-335.8	-48.9	-197.6	-72.8
18	-11.4	27.1	-171.2	-22.0	-148.4	-59.9
19	-14.2	27.7	-230.6	-69.6	-56.7	-17.5
20	-18.4	28.3	-1089.6	-308.1	-104.1	-43.7

the random effect factor using the strategy of Datta *et al.* (2011), that is, by testing for the presence of the random effects,  $v_i$ . At  $\alpha = 0.05$  level of significance, the test statistic (see Subsection 3.2)  $T = 24.3$ , while the critical value of  $\chi_{19}^2$  is 30.1. Thus, the null hypothesis that  $A = 0$  is not rejected. As a result,  $\hat{\theta}_i = x_i' \hat{\beta}$  is used as the estimate of  $\theta_i$ , according to Datta *et al.* (2011). However, the main issue is how to assess the uncertainty. We apply the three different methods investigated in Subsection 3.2 to this data, and obtain the square roots of the estimated MSPEs, denoted by DHM, BT, and MJ, respectively. Here the MSPE estimates are obtained by taking the exponentials of the corresponding log-MSPE estimates. The Monte-Carlo sample size for BT and MJ is  $K = 4000$ . The results are presented in Table 5. It is seen that the measures of uncertainty by DHM are always smaller than those by BT and MJ. This is not surprising because DHM does not take into account the potential variation in model selection. As for the comparison between BT and MJ, the latter measures are larger in most cases.

As another comparison, we also computed the standard EBLUPs (i.e., without testing the presence of the random effects) and their corresponding McJack estimates of  $\sqrt{\text{MSPE}}$ . The results are presented in the last two columns of Table 5, where  $\tilde{\theta}_i$  represents the EBLUPs and MJ the corresponding estimated  $\sqrt{\text{MSPE}}$ s. Note that the same data were also analyzed by Datta *et al.* (2011), who stated that, because the estimated MSPEs for DHM are much smaller than those for EBLUP, the DHM method is “significantly more accurate”. The results of our analysis show that this is not necessarily the case when additional variation in the model selection (by testing) is taken into account, and estimated correctly: Out of the 23 small areas, only 5 has smaller estimated  $\sqrt{\text{MSPE}}$  for DHM as compared to EBLUP when comparing the MJs for both (column 8 vs column 10).

Finally, there is one area, #5, for which all of the uncertainty measures give essentially the same results, 0.047 (although, to the fourth digit, the DHM measure is still smaller than its BT and MJ counterparts). This case corresponds to the “outlier” for this data, according to Jiang, Nguyen and Rao (2011). As noted by the latter authors (also see Jiang *et al.* 2010), without this case, a quadratic, instead of cubic, mean function would fit the data well. However, there is an over-fitting problem for this particular area, that is, the



Table 3: The Hospital Data, Estimates, and Measures of Uncertainty

Area	$y_i$	$s_i$	$\sqrt{D_i}$	$\hat{\theta}_i$	DHM	BT	MJ	$\tilde{\theta}_i$	MJ
1	.302	.112	.055	.221	.015	.029	.038	.238	.034
2	.140	.206	.053	.186	.013	.027	.019	.178	.019
3	.203	.104	.052	.214	.014	.029	.038	.215	.036
4	.333	.168	.052	.215	.011	.028	.044	.240	.040
5	.347	.337	.047	.349	.047	.047	.047	.349	.047
6	.216	.169	.046	.215	.011	.026	.030	.218	.024
7	.156	.211	.046	.183	.015	.027	.026	.176	.021
8	.143	.195	.046	.195	.011	.026	.032	.184	.034
9	.220	.221	.044	.177	.018	.029	.040	.186	.040
10	.205	.077	.044	.168	.015	.029	.048	.177	.049
11	.209	.195	.042	.195	.011	.026	.030	.199	.027
12	.266	.185	.041	.203	.010	.026	.029	.221	.026
13	.240	.202	.041	.189	.012	.026	.030	.203	.030
14	.262	.108	.036	.218	.014	.026	.021	.235	.018
15	.144	.204	.036	.188	.013	.025	.028	.174	.026
16	.116	.072	.035	.155	.017	.028	.038	.141	.042
17	.201	.142	.033	.228	.015	.025	.025	.221	.025
18	.212	.136	.032	.229	.015	.025	.025	.226	.025
19	.189	.172	.031	.213	.010	.023	.017	.205	.019
20	.212	.202	.029	.189	.012	.024	.038	.199	.034
21	.166	.087	.029	.189	.013	.024	.036	.180	.030
22	.173	.177	.027	.209	.010	.023	.032	.194	.034
23	.165	.072	.025	.155	.017	.022	.022	.159	.019

outlier causes the cubic fit to be “perfect” for this area. This means that the fitted cubic function goes through exactly the data point; as a result, the direct estimate,  $y_5$ , is equal to the regression estimate,  $x_5' \hat{\beta}$ . As a result, there is no difference between the EBLUP and the direct and synthetic estimates, regardless of the value of  $D_5$  and how one estimates  $A$ . Thus, in this case, every method essentially reduces to the direct estimate,  $y_5 = 0.347$ , and its measure of uncertainty,  $\sqrt{D_5} = 0.047$ .

Another real-data example on estimation of median income of four-person families is also considered. Again, the details are deferred to Supplementary Material.

## 6 Discussion

We have shown that the impact of model selection in accuracy measures may be complicated. If the accuracy measure only focuses on the variance, model selection is likely to add additional variation to the measure. This is shown, for example, in Subsection 3.1, where the EBLUP is an unbiased estimator, hence the MSPE reduces to the variance. On the other hand, if the accuracy measure is the MSPE, the overall impact of model selection depends on the relative contributions of the bias and variance as in the identity  $\text{MSPE} = (\text{prediction bias})^2 + \text{prediction variance}$ . As further discussed in Supplementary Material, model selection helps to reduce the bias but this may be at the cost of adding more variation. Because, in practice, it is difficult to predict in which way, and how much, the overall impact is, the best strategy is to obtain an accurate MSPE estimator. We have shown that the latter can be done via McJack.

## 7 Proofs

### 7.1 Proof of Theorem 1

Throughout this proof,  $\psi$  denotes the true parameter vector. Let  $\widetilde{b(\psi)}$  denote (15) with  $\widetilde{b}(\cdot)$  replaced by  $b(\cdot)$ . Also,  $c$  denotes a positive, generic constant, whose value may be

different at different places. By Theorem 5.2 of Jiang *et al.* (2002), we have

$$\mathbb{E}_y\{\widehat{b(\psi)} - b(\psi)\} = o(m^{-1-\gamma}), \quad (20)$$

where  $\gamma = [(d-2)/(2d+1)] \wedge \nu > 0$ , and  $\mathbb{E}_y$  denotes expectation with respect to  $y$ . Because the left side of (20) does not depend on  $\xi$ , the equation also holds with  $\mathbb{E}_y$  replaced by  $\mathbb{E}$ .

Let  $\mathbb{E}_\xi$  and  $\mathbb{P}_\xi$  denote expectation and probability with respect to  $\xi$ . Consider

$$\widehat{b(\psi)} - \widehat{b(\tilde{\psi})} = \hat{b}(\hat{\psi}) - b(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^m \{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j}) + b(\hat{\psi}) - \hat{b}(\hat{\psi})\}. \quad (21)$$

Let  $\tilde{\psi}$  be a fixed parameter vector such that  $|\tilde{\psi} - \psi| \leq w$ . Then, we have

$$\begin{aligned} \hat{b}(\tilde{\psi}) - b(\tilde{\psi}) &= \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} + \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(\tilde{s}(\tilde{\psi}) < c_1/2)} \\ &\quad + \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(\tilde{s}(\tilde{\psi}) > 2c_2)} \\ &= I_1 + I_2 + I_3. \end{aligned} \quad (22)$$

First note that, by A8, we have  $\mathbb{P}_\xi\{\tilde{s}(\tilde{\psi}) < c_1/2\} \leq \mathbb{P}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})| > c_1/2\} \leq (c_1/2)^{-q/2} \mathbb{E}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^{q/2}\}$ . Next, write  $u_k = \{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\}^2$  and note that  $\mathbb{E}_\xi(u_1) = s(\tilde{\psi})$ . By Marcinkiewicz-Zygmund inequality (e.g., Jiang 2010, p. 150), we have

$$\begin{aligned} \mathbb{E}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^{q/2}\} &= \frac{1}{K^{q/2}} \mathbb{E}_\xi \left[ \left| \sum_{i=1}^K \{u_k - \mathbb{E}_\xi(u_1)\} \right|^{q/2} \right] \\ &\leq \frac{c}{K^{q/2}} \mathbb{E}_\xi \left[ \sum_{k=1}^K \{u_k - \mathbb{E}_\xi(u_1)\}^2 \right]^{q/4} \\ &\leq \frac{c}{K^{q/4}} \times \frac{1}{K} \sum_{k=1}^K \mathbb{E}_\xi[|u_k - \mathbb{E}_\xi(u_1)|^{q/2}] \\ &\leq \frac{c}{K^{q/4}}, \end{aligned}$$

using Jensen's inequality for the second-to-last step, and A8 for the last step. It follows, by A8 and the definition of  $\hat{b}(\cdot)$ ,  $b(\cdot)$  that

$$|\mathbb{E}_\xi(I_2)| \leq cm^\rho K^{-q/4}. \quad (23)$$

By essentially the same argument, we also have

$$|\mathbb{E}_\xi(I_3)| \leq cm^\rho K^{-q/4}. \quad (24)$$

Now suppose that  $c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2$ . We also know that  $c_1 \leq s(\tilde{\psi}) \leq c_2$  by A8. Thus, for sufficiently large  $m$ , we have  $\hat{b}(\tilde{\psi}) = \tilde{b}(\tilde{\psi})$ . By Taylor series expansion, we have

$$\begin{aligned} \hat{b}(\tilde{\psi}) - b(\tilde{\psi}) &= \tilde{b}(\tilde{\psi}) - b(\tilde{\psi}) \\ &= \log\{\tilde{s}(\tilde{\psi})\} - \log\{s(\tilde{\psi})\} \\ &= \frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} - \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^2}{2s(\tilde{\psi})^2} + \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^3}{3\eta^3}, \end{aligned} \quad (25)$$

where  $\eta$  lies between  $s(\tilde{\psi})$  and  $\tilde{s}(\tilde{\psi})$ ; hence, we have  $\eta \geq c_1/2$ . It follows that

$$\left| \mathbb{E}_\xi \left[ \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^3}{3\eta^3} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| \leq \frac{8}{3c_1^3} \mathbb{E}_\xi \{ |\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^3 \} \leq cK^{-3/2}, \quad (26)$$

using an earlier inequality. Similarly, we have

$$\left| \mathbb{E}_\xi \left[ \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^2}{2s(\tilde{\psi})^2} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| \leq cK^{-1}. \quad (27)$$

Furthermore, note that  $\mathbb{E}_\xi\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\} = 0$ , thus, we have

$$\begin{aligned} \left| \mathbb{E}_\xi \left[ \frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| &= \frac{1}{s(\tilde{\psi})} \left| \mathbb{E}_\xi [\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2 \text{ or } \tilde{s}(\tilde{\psi}) > 2c_2)}] \right| \\ &\leq \frac{\mathbb{E}_\xi[\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2)}]}{s(\tilde{\psi})} \\ &\quad + \frac{\mathbb{E}_\xi[\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) > 2c_2)}]}{s(\tilde{\psi})}. \end{aligned}$$

By Hölder and Jensen's inequalities, A8 and an earlier result, we have

$$\begin{aligned} &\mathbb{E}_\xi[\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2)}] \\ &\leq [\mathbb{E}_\xi\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\}^{q/2}]^{2/q} [\mathbb{P}_\xi\{\tilde{s}(\tilde{\psi}) < c_1/2\}]^{1-2/q} \\ &\leq c \left\{ \frac{1}{K} \sum_{k=1}^K \mathbb{E}_\xi(u_k^{q/2}) + c_2^{q/2} \right\}^{2/q} K^{-(q/4)(1-2/q)} \\ &\leq cK^{-(q-2)/4}. \end{aligned}$$

Similarly, we have  $E_\xi[\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\}1_{(\tilde{s}(\tilde{\psi}) > 2c_2)}] \leq cK^{-(q-2)/4}$ . It follows that

$$\left| E_\xi \left[ \frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| \leq cK^{-(q-2)/4}. \quad (28)$$

Combining (24)–(28), and the fact that  $(q-2)/4 \geq 1$  by A8, we conclude that

$$|E_\xi(I_1)| \leq cK^{-1}. \quad (29)$$

Thus, combining (22)–(24), and (29), we have

$$|E_\xi\{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}| \leq c [m^\rho K^{-q/4} + K^{-1}], \text{ if } |\tilde{\psi} - \psi| \leq w, \quad (30)$$

where  $c$  does not depend on  $\tilde{\psi}$ .

Now, for any  $0 \leq j \leq m$ , we have

$$E\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\} = E_y[E_\xi\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})|\hat{\psi}_{-j}\}] = E_y\{\Delta(\hat{\psi}_{-j})\},$$

where  $\Delta(\tilde{\psi}) = E_\xi\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})|\hat{\psi}_{-j} = \tilde{\psi}\} = E_\xi\{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}$  by A7. Thus, we have

$$E\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\} = E_y\{\Delta(\hat{\psi}_{-j})1_{(|\hat{\psi}_{-j} - \psi| \leq w)}\} + E_y\{\Delta(\hat{\psi}_{-j})1_{(|\hat{\psi}_{-j} - \psi| > w)}\}. \quad (31)$$

By (30), the first term on the right side of (31) is bounded in absolute value by  $c[m^\rho K^{-q/4} + K^{-1}]$ . As for the second term, by the definition of  $\hat{b}(\cdot)$ ,  $b(\cdot)$ , and A5, it is bounded in absolute value by  $cm^{\rho-d}$ . Thus, in conclusion, we have

$$|E\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\}| \leq c [m^\rho K^{-q/4} + K^{-1} + m^{\rho-d}], \quad 0 \leq j \leq m. \quad (32)$$

Combining (21), (32), we have

$$|E\{\widehat{b(\psi)} - \widetilde{b(\psi)}\}| \leq c \left( m^{1+\rho} K^{-q/4} + \frac{m}{K} + m^{1+\rho-d} \right) = o(m^{-1}), \quad (33)$$

by A9 and the conditions on  $d, q$ .

The result then follows by (20) (with  $E_y$  replaced by  $E$ ) and (33).

## 7.2 Proof of Theorem 2

First, by (i)–(iv) of Jiang *et al.* (2002, p. 1803), it is easy to see that assumptions A1–A6 are satisfied. Assumption A7 is satisfied by the statement above A1. Thus, all we need is to verify assumption A8. Once again, in the arguments below,  $c$  denotes a positive constant whose value may be different at different places.

Suppose that the data are generated under the parameter vector  $\tilde{\psi}$ . Let  $\tilde{\theta}_i$  denote the BP of  $\theta_i$ . Then, we have  $s(\tilde{\psi}) = \text{MSPE}_{\tilde{\psi}}(\hat{\theta}_i) = \text{MSPE}_{\tilde{\psi}}(\tilde{\theta}_i) + \text{E}_{\tilde{\psi}}\{(\hat{\theta}_i - \tilde{\theta}_i)^2\} \geq \text{MSPE}_{\tilde{\psi}}(\tilde{\theta}_i) = \tilde{A}D_i/(\tilde{A} + D_i)$ . Thus, if  $0 < A/2 \leq \tilde{A} \leq 2A$ , where  $\tilde{A}$  is the  $A$  component of  $\tilde{\psi}$ , and  $A$  is the true  $A$ ,  $s(\tilde{\psi})$  is clearly bounded away from zero.

On the other hand, we have  $s(\tilde{\psi}) = \text{E}_{\tilde{\psi}}(\hat{\theta}_i^2) - 2\text{E}_{\tilde{\psi}}(\hat{\theta}_i\theta_i) + \text{E}_{\tilde{\psi}}(\theta_i^2)$ . By (8), we have  $\text{E}(\theta_i^2) \leq 2(|x_{f,i}|^2|\tilde{\beta}_f|^2 + \tilde{A}^2) \leq c$ , if, say  $|\tilde{\beta}_f - \beta_f| \leq 1$  and  $\tilde{A} \leq 2A$ . Also, by (11), and Jensen's inequality, we have

$$\hat{\theta}_i^2 \leq \frac{\hat{A}_f}{\hat{A}_f + D_i} y_i^2 + \frac{D_i}{\hat{A}_f + D_i} |x_{f,i}\hat{\beta}_f|^2 \leq y_i^2 + |x_{f,i}|^2 |\hat{\beta}_f|^2, \quad (34)$$

and, by (7),  $\text{E}_{\tilde{\psi}}(y_i^2) = \{\text{E}_{\tilde{\psi}}(y_i)\}^2 + \text{var}_{\tilde{\psi}}(y_i) = (x_{f,i}\tilde{\beta}_f)^2 + \tilde{A} + D_i \leq |x_{f,i}|^2 |\tilde{\beta}_f|^2 + \tilde{A} + D_i \leq c$ . Define  $P_f = I_m - D^{-1/2}X_f(X_f'D^{-1}X_f)^{-1}X_f'D^{-1/2}$ . By Lemma 1 of Jiang (2000), with  $\hat{V} = \hat{A}I_m + D$ ,  $D = \text{diag}(D_i, 1 \leq i \leq m)$ ,  $X_D = D^{-1/2}X_f$ ,  $Z = D^{-1/2}$ ,  $\Gamma = \hat{A}I_m$ , and  $\zeta = y - X_f\tilde{\beta}_f$ , we have

$$\begin{aligned} \hat{\beta}_f &= (X_f'\hat{V}^{-1}X_f)^{-1}X_f'\hat{V}^{-1}y \\ &= \tilde{\beta}_f + (X_f'\hat{V}^{-1}X_f)^{-1}X_f'\hat{V}^{-1}\zeta \\ &= \tilde{\beta}_f + \{X_D'(I_m + Z\Gamma Z')^{-1}X_D\}^{-1}X_D'(I_m + Z\Gamma Z')^{-1}D^{-1/2}\zeta \\ &= \tilde{\beta}_f + (X_D'X_D)^{-1}X_D'\{I_m - \hat{A}D^{-1}P_f(I_m + \hat{A}P_fD^{-1}P_f)^{-1}\}D^{-1/2}\zeta \\ &= \tilde{\beta}_f + (X_f'D^{-1}X_f)^{-1}X_f'D^{-1}\zeta \\ &\quad - (X_f'D^{-1}X_f)^{-1}X_f'D^{-1}\hat{A}D^{-1/2}P_f(I_m + \hat{A}P_fD^{-1}P_f)^{-1}D^{-1/2}\zeta \\ &= \tilde{\beta}_f + I_1 - I_2. \end{aligned} \quad (35)$$

Note that  $E_{\tilde{\psi}}(\zeta\zeta') = \tilde{A}I_m + D \leq cD$  under the assumptions. Thus, we have

$$\begin{aligned}
E_{\tilde{\psi}}(|I_1|^2) &= E_{\tilde{\psi}} \left[ \text{tr} \{ (X_f' D^{-1} X_f)^{-1} X_f' D^{-1} \zeta \zeta' D^{-1} X_f (X_f' D^{-1} X_f)^{-1} \} \right] \\
&= \text{tr} \{ (X_f' D^{-1} X_f)^{-1} X_f' D^{-1} E_{\tilde{\psi}}(\zeta \zeta') D^{-1} X_f (X_f' D^{-1} X_f)^{-1} \} \\
&\leq c \text{tr} \{ (X_f' D^{-1} X_f)^{-1} \} \\
&\leq \frac{c}{m \lambda_{\min}(m^{-1} X_f' X_f)}.
\end{aligned} \tag{36}$$

Furthermore, we have

$$|I_2| \leq \| (X_f' D^{-1} X_f)^{-1} X_f' D^{-1} \| \cdot \| \hat{A} D^{-1/2} P_f (I_m + \hat{A} P_f D^{-1} P_f)^{-1} \| \cdot |D^{-1/2} \zeta|.$$

By a similar argument as above, we have

$$\| (X_f' D^{-1} X_f)^{-1} X_f' D^{-1} \|^2 \leq \frac{c}{m \lambda_{\min}(m^{-1} X_f' X_f)}.$$

Next, let  $\lambda_1 \geq \dots \geq \lambda_m \geq 0$  be the eigenvalues  $P_f D^{-1} P_f$ . Then, we have

$$\| \hat{A} D^{-1/2} P_f (I_m + \hat{A} P_f D^{-1} P_f)^{-1} \|^2 = \max_{1 \leq i \leq m} \frac{\hat{A}^2 \lambda_i}{(1 + \hat{A} \lambda_i)^2}.$$

If  $\hat{A} \lambda_i = 0$ , then  $\hat{A}^2 \lambda_i / (1 + \hat{A} \lambda_i)^2 = 0$ ; otherwise,  $\hat{A}^2 \lambda_i / (1 + \hat{A} \lambda_i)^2 \leq \hat{A}^2 \lambda_i / \hat{A}^2 \lambda_i^2 = 1 / \lambda_i$ .

It follows that

$$\max_{1 \leq i \leq m} \frac{\hat{A}^2 \lambda_i}{(1 + \hat{A} \lambda_i)^2} \leq \frac{1}{\lambda_r},$$

where  $r = \text{rank}(P_f D^{-1} P_f)$ . Because  $P_f D^{-1} P_f u = 0$  if and only if  $P_f u = 0$ , we have  $r = \text{rank}(P_f)$ , and  $P_f$  is a projection matrix, whose eigenvalues are 0 or 1. Also, because

$$P_f D^{-1} P_f \geq \frac{P_f^2}{\max_{1 \leq i \leq m} D_i} = \frac{P_f}{\max_{1 \leq i \leq m} D_i},$$

by a well-known eigenvalue inequality (e.g., DasGupta 2008, p. 669), we have

$$\lambda_r \geq \lambda_r \left( \frac{P_f}{\max_{1 \leq i \leq m} D_i} \right) = \frac{\lambda_r(P_f)}{\max_{1 \leq i \leq m} D_i} = \frac{1}{\max_{1 \leq i \leq m} D_i}.$$

Thus, in conclusion, we have

$$\| \hat{A} D^{-1/2} P_f (I_m + \hat{A} P_f D^{-1} P_f)^{-1} \|^2 \leq \max_{1 \leq i \leq m} D_i.$$



Finally, it is easy to show that  $E_{\tilde{\psi}}(|D^{-1/2}\zeta|^2) \leq cm$ . Thus, combining the results, we have

$$E_{\tilde{\psi}}(|I_2|^2) \leq \frac{c}{\lambda_{\min}(m^{-1}X_f'X_f)}. \quad (37)$$

The upper bound for  $s(\tilde{\psi})$  follows from (34)–(37).

The last part of A8 follows from the above arguments by noting that  $\theta_i^{(k)} = x_{f,i}'\tilde{\beta}_f + \tilde{A}^{1/2}\xi_i^{(k)}$ ,  $y_i = \theta_i^{(k)} + \sqrt{D_i}\eta_i^{(k)}$ , and  $\xi_i^{(k)}, \eta_i^{(k)}$  are  $N(0, 1)$  random variables.

**Acknowledgements.** The research of Jiming Jiang is partially supported by the NSF grant SES-1121794. The research of Thuan Nguyen is partially supported by the NSF grant SES-1118469. The research of Jiming Jiang and Thuan Nguyen are partially supported by the NIH grant R01-GM085205A1.

## References

- [1] Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki eds.), Akademiai Kiado, Budapest, 267-281.
- [2] Berk, R., Brown, L., and Zhao, L. (2010), Statistical inference after model selection, *J. Quant. Criminol.* 26, 217-236.
- [3] Chen, S. and Lahiri, P. (2011), On the estimation of mean squared prediction error in small area estimation, *Calcutta Statist. Assoc. Bull.* 63, 249-252.
- [4] Das, K., Jiang, J. and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818-840.
- [5] DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York.
- [6] Datta, G. S., Hall, P., and Mandal, A. (2011), Model selection by testing for the presence of small-area effects, and applications to area-level data, *J. Amer. Statist. Assoc.* 106, 361-374.

- [7] Datta, G. S., Lahiri, P., and Maiti, T. (2002), Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data, *J. Statist. Planning Inference* 102, 83-97.
- [8] Efron, B. (1992), Jackknife-after-bootstrap standard errors and influence functions (with discussion), *J. Roy. Statist. Soc. Ser. B* 54, 83-127.
- [9] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96, 1348-1360.
- [10] Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269-277.
- [11] Ganesh, N. (2009), Simultaneous credible intervals for small area estimation problems, *J. Multivariate Anal.* 100, 1610-1621.
- [12] Gershunskaya, J. B. and Dorfman, A. H. (2013), Calibration and evaluation of generalized variance functions, *Proceedings of Joint Statistical Meetings, Survey Methods Section*, 2655-2669.
- [13] Hall, P. and Maiti, T. (2006), On parametric bootstrap methods for small area prediction, *J. Roy. Statist. Soc. Ser. B* 68, 221-238.
- [14] Jiang, J. (2000), A matrix inequality and its statistical application, *Linear Algebra Appl.* 307, 131-144.
- [15] Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.
- [16] Jiang, J. (2010), *Large Sample Techniques for Statistics*, Springer, New York.
- [17] Jiang, J. (2014), The fence methods, in *Advances in Statistics*, Vol. 2014, 1-14, Hindawi Publishing Corp.

- [18] Jiang, J., Lahiri, P., and Wan, S.-M. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782-1810.
- [19] Jiang, J., Nguyen, T. and Rao, J. S. (2010), Fence method for nonparametric small area estimation, *Survey Methodology* 36, 3-11.
- [20] Jiang, J., Nguyen, T. and Rao, J. S. (2011), Best predictive small area estimation, *J. Amer. Statist. Assoc.* 106, 732-745.
- [21] Jiang, J. and Rao, J. S. (2003), Consistent procedures for mixed linear model selection, *Sankhya A* 65, 23-42.
- [22] Lahiri, P. and Rao, J. N. K. (1995), Robust estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 90, 758-766.
- [23] Lahiri, P. and Suntornchost, J. (2014), Variable selection for linear mixed models with applications to small area estimation, *Sankhya*, DOI 10.1007/s13571-015-0096-0.
- [24] Leeb, H. (2009), Conditional predictive inference after model selection, *Ann. Statist.* 37, 2838-2876.
- [25] Molina, I., Rao, J. N. K., and Datta, G. S. (2015), Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects, *Survey Methodology*, in press.
- [26] Morris, C. N. and Christiansen, C. L. (1995), Hierarchical models for ranking and for identifying extremes with applications, *Bayes Statistics* 5, Oxford Univ. Press.
- [27] Müller, S., Scealy, J. L., and Welsh, A. H. (2013), Model selection in linear mixed models, *Statist. Sci.* 28, 135-167.
- [28] Pang, Z., Lin, B. and Jiang, J. (2015), Regularization parameter selections with divergent and NP-dimensionality via bootstrapping, *Austral. New Zealand J. Statist.*, in press.

- [29] Pfeiffermann, D. (2013), New important developments in small area estimation, *Statist. Sci.* 28, 40-68.
- [30] Prasad, N. G. N. and Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163-171.
- [31] Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, 2nd ed., Wiley, New York.
- [32] Rao, J. N. K. and Yu, M. (1994), Small area estimation by combining time series and cross-sectional data, *Canad. J. Statist.* 22, 511-528.
- [33] Rao, C. R. and Wu, Y. (2001), On model selection, in *IMS Lecture Notes–Monograph Series* 38, 1-57.
- [34] Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461-464.
- [35] Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B* 16, 385-395.
- [36] Yoshimori, M. and Lahiri, P. (2014), A second-order efficient empirical Bayes confidence interval, *Ann. Statist.* 42, 1233-1261.
- [37] Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M. P. (1999), An experimental comparison of ordinary and universal Kriging and inverse distance weighting, *Math. Geol.* 31, 375-390.